

**The Tang-Song Transition as a Textual Hypothesis:
Institutional Sequencing in Chinese Dynastic Chronicles**

September 2026
ISS Discussion Paper Series
F-204

Asei Ito^{*}
Jaehwan Lim[†]
Hongyong Zhang[‡]

^{*} Institute of Social Science, The University of Tokyo, Japan. asei@iss.u-tokyo.ac.jp

[†] School of International Politics, Economics, and Communication, Aoyama Gakuin University, Tokyo, Japan. t13025@aoyamagakuin.jp.

[‡] Research Group, The Research Institute of Economy, Trade, and Industry, Tokyo, Japan. zhang-hong-yong@rieti.go.jp.

The Tang-Song Transition as a Textual Hypothesis
Institutional Sequencing in Chinese Dynastic Chronicles

Asei Ito

Institute of Social Science, The University of Tokyo, Tokyo, Japan

Jaehwan Lim

School of International Politics, Economics, and Communication, Aoyama Gakuin University, Tokyo, Japan

Hongyong Zhang

Research Group, The Research Institute of Economy, Trade, and Industry, Tokyo, Japan

Abstract

Purpose: This paper asks whether classic historiographical interpretations of the Tang-Song transition can be operationalized as explicit expectations about the timing and sequence of institutional change in chronological histories.

Design/methodology/approach: Annual text for 618-1099 from the Zizhi Tongjian and Xu Zizhi Tongjian Changbian is divided into fixed-length chunks and represented with character 2-4 grams, Term Frequency-Inverse Document Frequency (TF-IDF), and a 50-dimensional truncated-Singular Value Decomposition (SVD) semantic space. Chunk vectors are averaged within calendar years. Event-concentrated n-grams are mechanically filtered, without an institutional whitelist, for a semantic-validation exercise. The main analysis maps historically derived institutional stages onto transparent institutional dictionaries.

Findings: The corpus displays an ordered institutional sequence but a mixed result for the classical elite-decline proposition. Regional military vocabulary rises from the eighth century and later declines; central military-administrative vocabulary rises during the Five Dynasties; and examination recruitment, local civil administration, and fiscal-commercial administration become more prominent later, primarily around and after the Song transition. A simple aristocratic-status dictionary is sparse and does not show a clean secular decline. Because 960 is also a source boundary, the pre-960 military sequence provides the cleaner evidence.

Originality: Rather than treating computational change points as historical discoveries, the paper begins with qualitative historiographical interpretations and translates them into explicit, observable textual expectations.

Contribution to the field of Digital Humanities: The study demonstrates a transparent workflow from historiographical interpretation to explicit hypotheses, textual proxies, corpus-wide measurement, and source-critical interpretation.

Keywords: Tang-Song transition; digital humanities; historiography; text as data; Zizhi Tongjian; institutional change; computational history

1. Introduction

The transition from Tang to Song is conventionally understood not simply as a dynastic change, but as a transition in the political and social organization of imperial China. In the classical formulation associated with Naitō (1914, 1922), its political core lay in the decline of aristocratic government and the emergence of a more centralized monarchical order, accompanied by a broader change in the social basis of officialdom. Miyazaki (1950, pp. 8-83) extended this interpretation into a broader account of Song “early modernity,” centered on the expansion of urban and commercial society and the rise of an examination-based bureaucratic elite. These political, social, and economic strands continue to structure modern accounts of the Tang-Song transition (Elvin 1973; Hartman 2015, pp. 19–24; Tackett 2017; von Glahn 2022).

Yet the Tang-Song transition is not a single event with an uncontested date or a single institutional dimension. Later scholarship has both extended its chronology and disaggregated its components. Hartwell (1982) situates political, demographic, and social change within a longer transition, while research on Tang elites and the Five Dynasties points to important continuities as well as substantial institutional changes before the founding of the Song in 960 (Tackett 2017, 2020; Hartman 2015, pp. 19–24). Wang (2022) more recently recasts the Tang-Song transition as a shift in the elite social relations, from a centralized network toward more localized and fragmented networks, strengthening monarchical power over the elite. The historiographical problem is therefore better understood as one of sequence and timing across multiple dimensions of change than as the identification of a single break point.

This historiographical structure poses a natural problem for computational history. If the Tang-Song transition consisted of a sequence of changes rather than a single break, can those historically specified changes be

translated into observable expectations in a long chronological corpus? Recent quantitative-history scholarship has demonstrated the potential of systematic data construction for studying long-run Chinese state and social development (Chen and Ma 2023; Chen, Campbell, and Ma 2026). The present paper takes a complementary text-as-data approach. Rather than inferring historical transition from the mere existence of computational change points—which a sufficiently long corpus will almost inevitably contain—we begin with historiographical interpretation, derive observable textual expectations in advance, and assess them across the full chronological corpus.

The digital-humanities contribution lies not in any single computational technique, but in the operationalization workflow that links historiography to corpus measurement. We begin with a historiographical interpretation, derive explicit expectations from it, map those expectations onto transparent textual proxies, measure the proxies across the full corpus, and then return to source criticism to interpret the resulting temporal patterns. This movement from concepts to texts and back parallels Pichler and Reiter’s (2022) account of operationalization as a core activity of digital humanities. We therefore do not claim that Term Frequency-Inverse Document Frequency (TF-IDF), Singular Value Decomposition (SVD), or change-point detection independently discovers a historical periodization. Rather, computational methods allow us to assess systematically whether a historically derived sequence of institutional change is recoverable from the chronological textual record. This approach also aligns with digital-history scholarship emphasizing that computational scale complements rather than substitutes for historical interpretation and source criticism (Gregory 2014; Robertson 2022).

The empirical strategy has two steps. Hypothesis 1 (H1) is a validation exercise: after short-lived wars, actors, and other event-specific language are reduced, sustained semantic movement should remain across the late Tang, Five Dynasties, and early Song. Hypothesis 2 (H2) is the main historical assessment: it translates historically derived expectations about elite and institutional change into transparent dictionary-based measures and evaluates their temporal ordering. Because the source changes at 960, pre-960 movements receive greater evidentiary weight.

2. Literature, Operationalization, and Hypotheses

For operationalization, two qualifications in the historical literature are especially important. First, the Tang-Song transition is better understood as an extended and multidimensional process than as a single break at the dynastic

boundary of 960. Tackett (2017) frames this transformation as a series of changes extending from the mid-Tang to the mid-Southern Song, encompassing changes in the sociopolitical elite, political institutions and culture, and the economy, rather than a simultaneous rupture at 960. Von Glahn (2022, pp. 246–250) identifies a more specific political-economic sequence from the An Lushan Rebellion (755–763) and the fiscal reorganization culminating in the twice-a-year tax of 780, through late-Tang regional fragmentation and the Five Dynasties, to Song consolidation. Second, recent prosopographic work complicates a simple model of steadily declining Tang aristocracy: old aristocratic families remained deeply embedded in the political elite through the late Tang (Tackett 2020, pp. 277–279). We therefore retain the declining political centrality of the aristocratic order as an *ex ante* historical expectation, while treating aristocratic-status vocabulary as a deliberately weak textual proxy for that expectation. The paper focuses on political-administrative and elite dimensions that can be measured transparently in chronological histories, rather than attempting to capture the full economic transformation described by Miyazaki (1950) and Elvin (1973).

On this basis, we formulate two hypotheses. H1 provides an unsupervised validation of long-run semantic change, while H2 specifies a set of *ex ante* expectations about elite and institutional sequencing derived from the historical literature.

H1 (validation: long-run semantic change): If the Tang-Song transition involved sustained political and administrative reorganization, event-filtered semantic change should remain visible across the late Tang-Five Dynasties-early Song period after short-lived event language is reduced.

H2 (main assessment: elite and institutional sequencing): The historical literature implies an ordered institutional transition rather than a single break at 960. We therefore specify six *ex ante* observable predictions:

H2a (aristocratic elite): Aristocratic-status vocabulary should become less prominent over the Tang-Song transition, reflecting the declining political centrality of the old aristocratic order (Naitō 1922). This expectation, however, is weaker than other institutional expectations: established aristocratic families remained prominent through the late Tang (Tackett 2020, pp. 277–279), and recent tomb-epitaph evidence shows that the declining career advantage of aristocratic pedigree did not entail a decline in the prevalence of prominent-branch status (Wen, Wang, and Hout 2024).

H2b (regional military governance): Regional military-governance vocabulary should rise from the mid-eighth century following the An Lushan Rebellion and later retreat with political consolidation (Hartwell 1982, pp. 395–397; von Glahn 2022, pp. 246–250).

H2c (central military administration): Central military-administrative vocabulary should rise later, particularly during the Five Dynasties, as competing regimes developed institutions for consolidating military and fiscal authority (von Glahn 2022, pp. 247–250).

H2d (examination recruitment): Examination-recruitment vocabulary should become increasingly prominent over a longer interval from the Tang into the Song, reflecting the growing role of competitive examinations in elite recruitment. Tomb-epitaph evidence indicates that examination success became increasingly important for bureaucratic advancement already within the Tang (Wen, Wang, and Hout 2024), while by the Song examinations had become a major route into the civil service and conferred substantial career advantages (Hartman 2015, pp. 55–58; Tackett 2017, pp. 118–128).

H2e (local civil administration): Local civil-administrative vocabulary should become more prominent with Song consolidation, reflecting the reorganization and functional differentiation of territorial administration and the strengthening of centrally supervised civil offices at the local level (Hartman 2015, pp. 51–52).

H2f (fiscal-commercial administration): Fiscal-commercial administrative vocabulary should become more prominent with Song consolidation, reflecting the expansion, functional differentiation, and centralization of fiscal administration, commercial taxation, and state financial administration (Hartwell 1982, pp. 397–398; Hartman 2015, pp. 22–24).

There is an important dependence between historiography and source. The scholarship from which H2 is derived itself draws partly on dynastic and chronological histories related to the texts analysed here. The exercise is therefore not a fully independent test of the Tang-Song thesis. Its contribution lies instead in asking whether interpretations developed through conventional historical scholarship are recoverable as systematic temporal patterns under explicit and reproducible measurement rules. This dependence makes source criticism part of the computational design rather than an afterthought.

3. Data and Measurement

3.1 Chronological histories

The corpus is constructed from digitized texts obtained from the Kanripo (Kanseki Repository) maintained by the Institute for Research in Humanities, Kyoto University: the main text of the Zizhi Tongjian for 618-959 and the Xu Zizhi Tongjian Changbian for 960-1099. Editorial notes and paratext are removed, and the remaining text is assigned to calendar years. The resulting data contain year, source, reign-era information, annual text, and annual character counts. The 618–1099 analysis interval contains 477 observed annual text records across 482 calendar years ($477/482=99.0\%$). Years without a separable annual record are retained as missing rather than coded as zero-text years: 1068, 1069, and 1094–1096 are missing from the current derived Xu Zizhi Tongjian Changbian corpus. Descriptive statistics for the analysis corpus, including annual text volume, are reported in Appendix A. The change of source at 960 is the principal identification limitation. Changes observed before 960—especially the shift from regional military to central military-administrative vocabulary—are within the Zizhi Tongjian and cannot be mechanically generated by the source boundary. Changes clustered at or after 960 are treated as supportive but less clean evidence.

3.2 Event-filtered semantic validation

The semantic analysis follows a latent-semantic-analysis pipeline based on TF-IDF and truncated SVD (Deerwester et al. 1990), using character n-grams to avoid dependence on modern word segmentation of Classical Chinese (Tian and Kübler 2021; Wang 2020). Annual text is divided into consecutive 500-character chunks, with final fragments shorter than 250 characters discarded. Chunks are represented by character 2–4 grams; features appearing in fewer than five chunks are removed. The resulting chunk-feature matrix is TF-IDF weighted and reduced to a 50-dimensional semantic space, and chunk vectors are averaged within calendar years to obtain annual semantic vectors. Using chunks rather than years as the document unit limits the influence of unusually long annual entries.

Long-run semantic shift is measured as cosine distance between 30-year rolling semantic centroids separated by a 20-year lag. Let $\mathbf{v}_t$ denote the annual 50-dimensional semantic vector obtained by averaging the chunk vectors assigned to calendar year t , and let $\mathbf{v}_{t,30}$ denote the normalized mean of the annual semantic vectors over a

30-year rolling window, implemented as years $t-14$ through $t+15$. We define the Long-run Semantic Shift Index (LSSI) as follows:

$$LSSI_t = 1 - \cos(\mathbf{v}_{t,30}, \mathbf{v}_{t-20,30})$$

Higher values indicate greater sustained change in the overall semantic composition of the chronicle relative to the comparison centroid twenty years earlier. The rolling windows are defined on the calendar-year scale. Missing years are neither assigned zero vectors nor interpolated. For each year t , the 30-year centroid is calculated from the observed annual semantic vectors falling within years $t-14$ through $t+15$, provided that at least 20 annual vectors are available in that window; the centroid is then compared with the centroid defined for calendar year $t-20$. Because each centroid aggregates approximately three decades of text and the two centroids overlap in calendar time, the index is intended to capture long-run movement rather than to date a discrete annual break.

To reduce short-lived event language, n -grams are filtered mechanically and independently of the H2 dictionaries. For each n -gram, we calculate the maximum share of its corpus occurrences falling within any consecutive ten-year window and remove it when this share is at least 0.70. TF-IDF and the 50-dimensional SVD are then re-estimated on the filtered feature matrix. No institutional whitelist is used, preserving H1 as an unsupervised validation exercise distinct from the dictionary-based H2 analysis. Alternative concentration thresholds and windows produce highly similar trajectories (Appendix A, Table A2).

3.3 Institutional dictionaries

Translating H2a–H2f into measurable text proxies requires a transparent and historically grounded vocabulary-selection procedure. Candidate terms are evaluated against three criteria before dictionary-level trajectories are examined. First, construct validity requires a term to directly denote the institutional concept specified by the corresponding sub-hypothesis, with a historiographical basis for that identification (Hartman 2015; Hartwell 1982; Tackett 2017). Second, lexical specificity excludes generic or polysemous strings whose matches are dominated by unrelated senses, assessed through keyword-in-context (KWIC) checks. Third, temporal interpretability requires a term to retain the same relationship to the target construct wherever it occurs, rather than primarily reflecting transient events, reign-specific labels, or unrelated historical usages. Terms are not excluded merely because the institutions they denote emerged during the study period, since such emergence may itself constitute

part of the hypothesized transition. Term frequency and first attested year are treated as diagnostics rather than selection criteria, and dictionary size is allowed to vary across domains.

Eligible terms are then divided by canonicity into base and expanded dictionaries. The base dictionary (Appendix B, Table B1), used for the main analysis, contains canonical and direct indicators of each domain, such as principal institutional or commissioner titles. The expanded dictionary additionally includes valid but more peripheral indicators, including broader synonyms, subordinate or parallel offices, variant institutions, and related bureau or system forms. It is used to test whether the inferred temporal patterns depend on the narrower canonical specification (Appendix B, Table B2).

KWIC checks also identify systematic homographs and segmentation false positives, which are excluded from the reported counts. Where dictionary terms overlap, counts use longest-match-first attribution to avoid double counting.

4. Results

4.1 Long-run semantic shifts

Before event filtering, the semantic index shows very large changes in the 850s and 860s. N-gram inspection shows that these peaks are dominated by changing conflicts, regions, and actors. After temporally concentrated event language is filtered, the dominant long-run movement shifts toward the late ninth and early tenth centuries, with additional movement around the early Song (Fig. 1). The filtered trajectory does not identify either 907 or 960 as a singular semantic break. Instead, elevated semantic movement extends across the late Tang and into the Five Dynasties, with major peaks on both sides of 907, while the movement around 960 is smaller than the largest pre-960 shifts. This pattern is more consistent with a prolonged reorganization spanning dynastic boundaries than with a transition concentrated at a single conventional date.

This supports H1 in a limited but substantively informative sense: sustained semantic movement remains after obvious event language is reduced, and its timing places substantial change before the Song founding and before the source boundary at 960. More importantly, the exercise illustrates a general methodological point for chronicle-as-data research: a computationally prominent pattern may describe changing narrative content without directly measuring the historical process of interest.

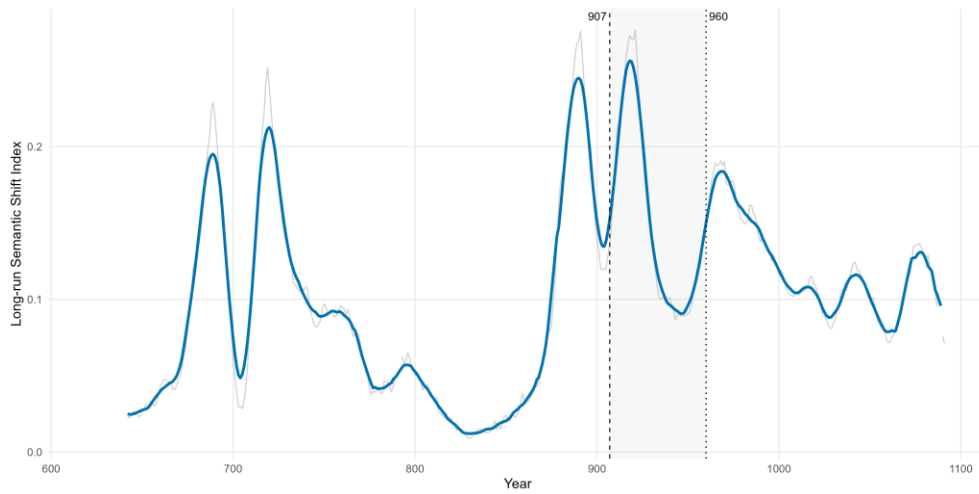


Fig. 1 Event-filtered long-run semantic shift

Notes: The horizontal axis is calendar year (618–1099); the vertical axis is the centered 11-year moving average of the Long-run Semantic Shift Index (LSSI), measured as cosine distance between event-filtered 30-year semantic centroids separated by 20 calendar years. The dashed line marks 907 (fall of the Tang); the dotted line marks 960 (founding of the Song and the corpus source boundary).

4.2 Institutional sequencing

H2 is assessed with the base dictionaries constructed in Section 3.3. Their empirical support differs substantially: examination-recruitment terms are frequent (e.g., 進士 *jìnshì*, 875 occurrences; 貢舉 *gòngjǔ*, 264), whereas most aristocratic-status terms are rare (e.g., 士族 *shìzú*, 10; 門閥 *ménfá*, 13). This contrast is itself informative about which historiographical claims can be readily operationalized in a chronicle corpus.

Table 1 provides a first summary of the temporal ordering across the six domains. Regional military-governance vocabulary rises sharply from the eighth century and reaches its highest period mean in 850–899, before declining thereafter. Central military-administrative vocabulary rises later, reaching its highest period mean in 900–949, while examination recruitment, local civil administration, and fiscal-commercial administration become substantially more prominent primarily from 950 onward. Importantly, the shift from regional to central military vocabulary is already visible before 960 and therefore occurs within the same source, although the later Song-period increases must be interpreted more cautiously because they coincide with the source boundary.

Appendix B reports the complete base dictionary and its base-versus-expanded robustness check, together with further robustness checks (leave-one-term-out and individual base-term trajectories).

Period	Aristocratic elite	Regional military	Central military	Examination recruitment	Local civil	Fiscal-commercial admin.
618–699	0.14	0.08	0.00	0.65	0.75	0.00
700–749	0.07	13.30	0.00	2.41	1.69	0.42
750–799	0.06	52.33	0.00	1.38	0.98	1.84
800–849	0.40	57.87	0.41	2.12	0.55	2.53
850–899	0.20	78.61	0.99	1.93	1.70	2.04
900–949	0.24	58.89	6.44	0.79	0.88	0.93
950–999	0.20	22.44	3.96	7.83	7.85	5.99
1000–1049	0.58	6.81	1.53	7.11	10.60	11.73
1050–1099	0.25	4.21	1.16	4.38	6.44	5.62

Table 1 Six-domain mean occurrences per 10,000 characters by period

Notes: Each category is the sum of component-term occurrences in a calendar year divided by that year's cleaned character count and multiplied by 10,000. Entries are arithmetic means of these annual normalized values within each period. The aristocratic-elite vocabulary is sparse and does not provide a validated measure of elite political power (also see Fig. B1).

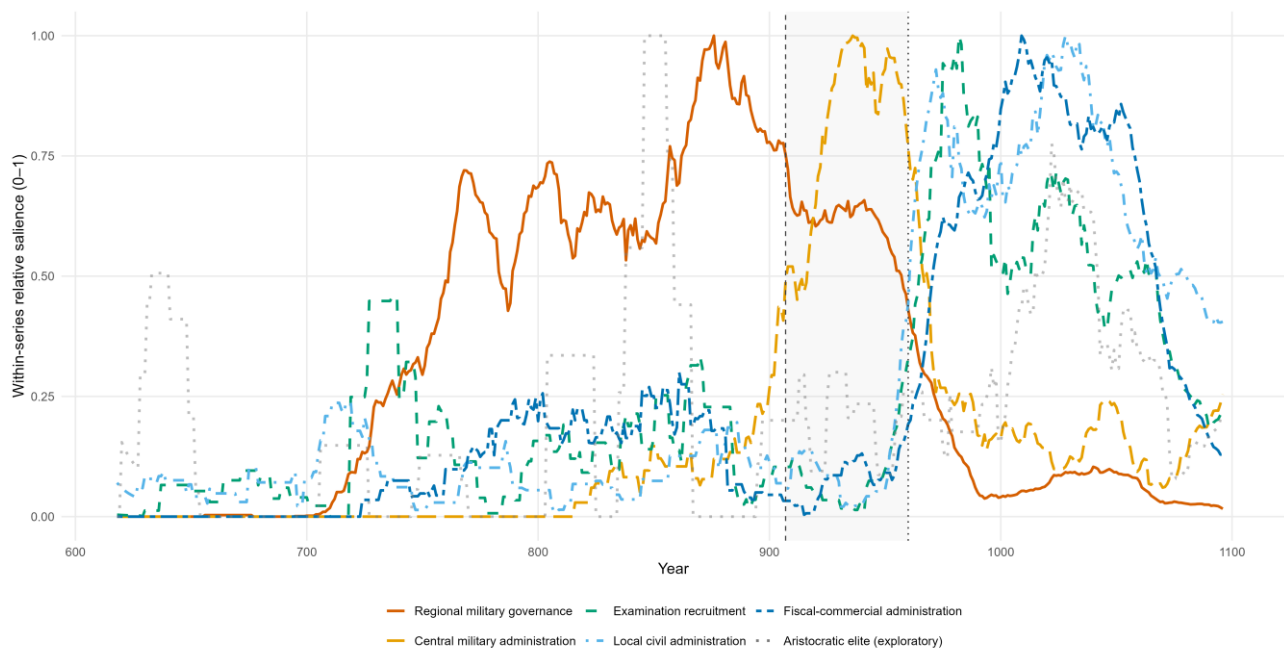


Fig. 2 Institutional-vocabulary sequencing, 618-1099

Notes: Each series shows the 21-year moving average of dictionary counts per 10,000 characters, normalized by its own sample maximum. The figure therefore compares the timing and trajectory, rather than the absolute

frequency, of institutional vocabularies. The dashed and dotted vertical lines mark 907 (end of the Tang) and 960 (founding of the Song), respectively.

Figure 2 yields a more differentiated result than a simple confirmation of the classical thesis. In the base specification, the predicted sequence is visible in the timing of the domain-specific peaks. Regional military governance rises from an early-Tang baseline, peaks in 850–899 (78.6 per 10,000 characters), and then declines through the Five Dynasties and Song. Central military-administrative vocabulary rises later and sharply, reaching its highest period mean in 900–949 (6.44), before receding under the Song. Examination recruitment and local civil administration rise most strongly after 950. Fiscal-commercial administration follows a more complex trajectory: in the base specification, it remains relatively low through the Tang and Five Dynasties before rising sharply after the Song founding and peaking in 1000–1049 (11.73). The base and expanded series are compared directly in Appendix B (Fig. B2).

Taken together, these results support an ordered but textured transition. Regional military governance rises first and later recedes; central military-administrative vocabulary becomes most prominent during the Five Dynasties; and examination recruitment, local civil administration, and fiscal-commercial administration become more prominent later, primarily from the early Song onward. H2b–H2f are broadly consistent with the predicted institutional sequence, although the fiscal-commercial trajectory is distinctly non-monotonic before its Song-period rise. The main ordering is generally preserved under the expanded dictionaries, although central military and fiscal-commercial administration are more sensitive to dictionary scope (Appendix B, Table B2, Fig. B2).

By contrast, aristocratic-status vocabulary does not show the predicted secular decline (H2a). The series is sparse and non-monotonic and remains broadly similar under the base and expanded dictionaries ($r \approx 0.93$). We therefore do not claim to recover Naitō's aristocratic-decline proposition from lexical counts alone; that dimension likely requires a clan- or prosopography-based measure capable of distinguishing the prevalence of aristocratic status from its changing political returns.

5. Discussion

The convergence between H1 and H2 strengthens the interpretation of the late Tang and Five Dynasties as a major transition period. H1 is an unsupervised, dictionary-free validation exercise, whereas H2 uses historiographically

derived institutional dictionaries constructed without reference to the semantic-shift timing. Despite these different measurement strategies, both locate substantial change in the late ninth and early tenth centuries: the event-filtered LSSI shows major movement around the Tang–Five Dynasties transition, while the institutional measures identify the late-Tang peak of regional military governance and the subsequent rise of central military-administrative vocabulary. This convergence is not an artifact of shared vocabulary: removing the forty-one base-dictionary terms from the underlying text before re-estimating the semantic space leaves the event-filtered LSSI trajectory essentially unchanged ($r = 0.99$), confirming that H1 and H2 provide complementary rather than circular evidence for an extended institutional transition rather than a single dynastic break.

The source boundary at 960 remains the principal limitation. The rise and retreat of regional military vocabulary (H2b), concentrated in the eighth and ninth centuries, provides the cleanest pre-960 evidence. Central military-administrative vocabulary (H2c) rises sharply in 900–949, indicating that its emergence predates the source change, although its subsequent trajectory cannot be fully separated from it. The later prominence of examination recruitment (H2d), local civil administration (H2e), and fiscal-commercial administration (H2f) is consistent with the hypothesized continuation of the sequence but is more entangled with the source boundary. The evidence therefore favors an ordered transition beginning before 960, while the post-960 components require greater caution.

Not every historiographical proposition is equally recoverable through lexical prevalence. The predicted decline in aristocratic-status vocabulary (H2a) is not recovered: the dictionary is sparse and non-monotonic, consistent with the possibility that lexical prevalence captures the continued salience of aristocratic identities rather than their declining political returns (Wen, Wang, and Hout 2024). This asymmetry is itself informative: institutional-administrative change is more readily represented in chronicle vocabulary than changing elite power. More broadly, because the historiography underlying H2 partly draws on the same historical record, the analysis should be understood as a transparent and reproducible operationalization of historical interpretations rather than an independent external test.

6. Conclusion

This paper asks whether classic interpretations of the Tang-Song transition can be recovered as systematic temporal patterns under explicit corpus-level measurement. The evidence is strongest for an ordered institutional

transition rather than a common break at 960: regional military governance rises and later recedes; central military-administrative vocabulary rises across the Five Dynasties before the source boundary; and examination recruitment, local civil administration, and fiscal-commercial administration become more prominent later. A simple aristocratic-status dictionary, however, does not recover the predicted decline of the old elite order. The substantive conclusion is therefore selective rather than total: the corpus reveals a strong sequence of changes in military governance, recruitment, and civil-fiscal administration, while declining aristocratic political centrality requires a richer prosopographic or clan-based operationalization. The digital-humanities contribution of this study lies in making explicit the distinction between claims that can and cannot be recovered from a given textual representation.

Data availability

The source texts are openly available through the Kanripo repository maintained by the Institute for Research in Humanities, Kyoto University (Zizhi Tongjian, KR2b0007; Xu Zizhi Tongjian Changbian, KR2b0019, licensed under CC BY-SA 4.0).

Funding

This work was supported by the Japan Society for the Promotion of Science KAKENHI grant numbers JP24K03158.

References

- Chen, Z. and Ma, C. (2023). ‘Quantitative history studies on China: state capacity, institutions, culture and human capital from prehistoric times to the present’. *Asia-Pacific Economic History Review*, 63(2): 128–44.
- Chen, Z., Campbell, C. and Ma, D. (eds) (2026). *Quantitative History of China: State Capacity, Institutions and Development*. Singapore: Springer.
- Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K. and Harshman, R. (1990). ‘Indexing by latent semantic analysis’. *Journal of the American Society for Information Science*, 41(6): 391–407.
- Elvin, M. (1973). *The Pattern of the Chinese Past*. Stanford, CA: Stanford University Press.

- Gregory, I. (2014). Challenges and opportunities for digital history. *Frontiers in Digital Humanities*, 1, 1.
- Hartman, C. (2015). 'Sung government and politics'. In Chaffee, J. W. and Twitchett, D. (eds), *The Cambridge History of China, Volume 5, Part Two: Sung China, 960–1279*. Cambridge: Cambridge University Press, pp. 19–138.
- Hartwell, R. M. (1982). 'Demographic, political, and social transformations of China, 750–1550'. *Harvard Journal of Asiatic Studies*, 42(2): 365–442.
- Hinrichs, U., Alex, B., Clifford, J., Watson, A., Quigley, A., Klein, E. and Coates, C. M. (2015). 'Trading consequences: a case study of combining text mining and visualization to facilitate document exploration'. *Digital Scholarship in the Humanities*, 30(Suppl. 1): i50–i75.
- Michel, J.-B., Shen, Y. K., Presser Aiden, A. et al. (2011). 'Quantitative analysis of culture using millions of digitized books'. *Science*, 331(6014): 176–82.
- Miyazaki, I. (1950). *Tōyōteki kinsei* [The Early Modern in East Asia]. Osaka: Kyōiku Taimususha.
- Naitō, K. (1914). *Shina Ron* [On China]. Tokyo: Bunkaidō Shoten.
- Naitō, K. (1922). 'Gaikatsuteki Tō-Sō jidai kan' [A Synoptic View of the Tang and Song Periods]. *Rekishi to Chiri*, 9(5): 1–12.
- Pichler, A., & Reiter, N. (2022). From concepts to texts and back: Operationalization as a core activity of digital humanities. *Journal of Cultural Analytics*, 7(4).
- Robertson, S. (2022). The properties of digital history. *History and Theory*, 61(4), 86-106.
- Tackett, N. (2017). 'A Tang–Song turning point'. In Szonyi, M. (ed.), *A Companion to Chinese History*. Hoboken, NJ: Wiley Blackwell, pp. 118–28.
- Tackett, N. (2020). 'The evolution of the Tang political elite and its marriage network'. *Journal of Chinese History*, 4: 277–304.
- Tian, Z. and Kübler, S. (2021). 'Period classification in Chinese historical texts'. In *Proceedings of the 5th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*. Association for Computational Linguistics, pp. 168–77.
- von Glahn, R. (2022). 'The Tang-Song transition in Chinese economic history'. In Ma, D. and von Glahn, R. (eds), *The Cambridge Economic History of China, Volume 1: To 1800*. Cambridge: Cambridge University Press, pp. 243–56.

Wang, Y. (2022). *The Rise and Fall of Imperial China: The Social Origins of State Development*. Princeton, NJ: Princeton University Press.

Wang, Y.-C. (2020). ‘Word segmentation for classical Chinese Buddhist literature’. *Journal of the Japanese Association for Digital Humanities*, 5(2): 154–72.

Wen, F., Wang, E. H. and Hout, M. (2024). ‘Social mobility in the Tang Dynasty as the imperial examination rose and aristocratic family pedigree declined, 618–907 CE’. *Proceedings of the National Academy of Sciences*, 121(4): e2305564121.

Appendix A. Corpus and Semantic Robustness

Source	Observed years	Coverage	Total characters	Mean/year	Median/year	Min	Max
Zizhi Tongjian	342 (342/342=100%)	618–959	1,013,273	2,963	2,170	241	18,530
Xu Zizhi Tongjian Changbian	135 (135/140=96.4%)	960–1099	4,147,184	30,720	16,851	2,688	332,746
Full corpus	477 (477/482=99.0%)	618–1099	5,160,457	10,819	3,483	241	332,746

Table A1 Descriptive statistics for the year-level analysis corpus

Notes: Character counts refer to cleaned annual text used in the analysis. The corpus contains 477 observed year records over 618–1099. Years without an annual record in the current derived dataset are 1068, 1069, and 1094–1096. Means and medians are calculated across observed annual records.

Specification	Correlation with baseline	Mean absolute difference	Maximum absolute difference
10y / 60%	0.966	0.009	0.081
10y / 70% (baseline as shown in Figure 1)	1.000	0.000	0.000
10y / 80%	0.997	0.004	0.018
5y / 70%	0.976	0.011	0.051

Table A2 Similarity of sensitivity LSSI series to the 10-year/70% baseline

Notes: Correlations use years observed in both series.

Appendix B. Institutional Dictionaries and Robustness

Category	Term	Pinyin	English gloss
Aristocratic elite	士族	shìzú	elite lineage
	世族	shìzú	hereditary prominent lineage
	門閥	ménfá	aristocratic pedigree/clan status
	甲族	jiǎzú	first-rank lineage
	名族	míngzú	distinguished lineage
	大族	dàzú	great lineage
	衣冠	yīguān	elite/gentry status
Regional military governance	節度使	jiédùshǐ	military governor
	留後	liúhòu	acting military governor/successor
	牙兵	yábīng	headquarters troops
	牙軍	yájūn	headquarters army
	兵馬留後	bīngmǎ liúhòu	acting commander of troops
	藩鎮	fānzhèn	regional military governorship (system)
Central military administration	樞密使	shūmìshǐ	military affairs commissioner
	樞密院	shūmìyuàn	Bureau/Council of Military Affairs

	殿前都	diànqián dū	Palace Command (title stem)
	殿前司	diànqián sī	Palace Command (bureau)
	侍衛親軍	shìwèi qīnjūn	Palace Guard Army
	三衙	sānyá	Three Commands (umbrella term)
Examination recruitment	進士	jìnshì	presented scholar degree-holder
	貢舉	gòngǔ	examination nomination/recruitment
	制科	zhìkē	special imperial examination
	明經	míngjīng	classics examination
	及第	jídì	passing an examination
	登第	dēngdì	attaining an examination degree
	登科	dēngkē	passing the examination
	科舉	kējǔ	civil service examinations
	殿試	diànshì	palace examination
	鄉貢	xiānggòng	locally nominated tribute candidate
	貢院	gòngyuàn	examination hall
Local civil administration	通判	tōngpàn	assistant prefect/controller-general
	知州	zhīzhōu	prefectural administrator

	知縣	zhīxiàn	county magistrate
	縣令	xiànlǐng	county magistrate (Tang usage)
	知府	zhīfǔ	prefect (superior prefecture)
	縣尉	xiànwèi	county sheriff/constable
Fiscal-commercial administration	轉運使	zhuǎnyùnshǐ	transport/fiscal commissioner
	鹽鐵使	yántiěshǐ	salt and iron commissioner
	度支使	dùzhīshǐ	finance/expenditure commissioner
	三司使	sānsīshǐ	commissioner of the Three Fiscal Agencies
	市舶使	shíbóshǐ	maritime trade commissioner

Table B1 Complete base dictionary used in Figure 2

Notes: Counts use exact string matching. The expanded dictionary additionally includes the following 44 terms:

aristocratic elite: 世家, 舊族, 搢紳, 縉紳, 甲第, 膏粱, 右族, 世胄, 華胄; regional military governance: 觀察使, 團練使, 防禦使, 牙將, 鎮將, 支郡, 方鎮, 押衙, 都押衙; central military administration: 都部署, 巡檢, 巡檢, 禁軍, 六軍, 兵馬都監, 兵部; examination recruitment: 省試, 解試, 童子科; local civil administration: 監司, 錄事參軍, 提刑, 縣丞, 郡守, 令佐, 提點刑獄; and fiscal-commercial administration: 度支, 鹽鐵, 市舶, 常平, 常平倉, 提舉常平, 商稅, 榷貨務, 榷茶. The expanded specification therefore contains 85 terms in total, including the 41 base terms listed in the table.

Domain	Correlation (base vs. expanded)	Mean absolute difference	Maximum absolute difference
Aristocratic elite	0.928	0.17	0.34
Regional military governance	0.997	7.07	15.57
Central military administration	0.773	5.93	9.95
Examination recruitment	1.000	0.04	0.14
Local civil administration	0.980	0.56	3.22
Fiscal-commercial administration	0.807	3.43	7.44

Table B2 Robustness of period-mean institutional-vocabulary trajectories to base vs. expanded dictionary choice

Domain	Term dropped (worst case)	Correlation vs. full base	Mean absolute difference	Maximum absolute difference
Aristocratic elite	衣冠	0.690	0.14	0.40
Regional military governance	節度使	0.964	27.40	63.12
Central military administration	樞密使	0.671	0.93	5.35
Examination recruitment	明經	0.991	0.25	1.18
Local civil administration	縣令	0.993	0.96	1.72
Fiscal-commercial administration	轉運使	0.932	2.66	9.43

Table B3 Leave-one-term-out robustness: worst-case single-term dependency per domain (base dictionary)

Note: For each domain, the single base-dictionary term whose removal most disrupts the period-mean trajectory.

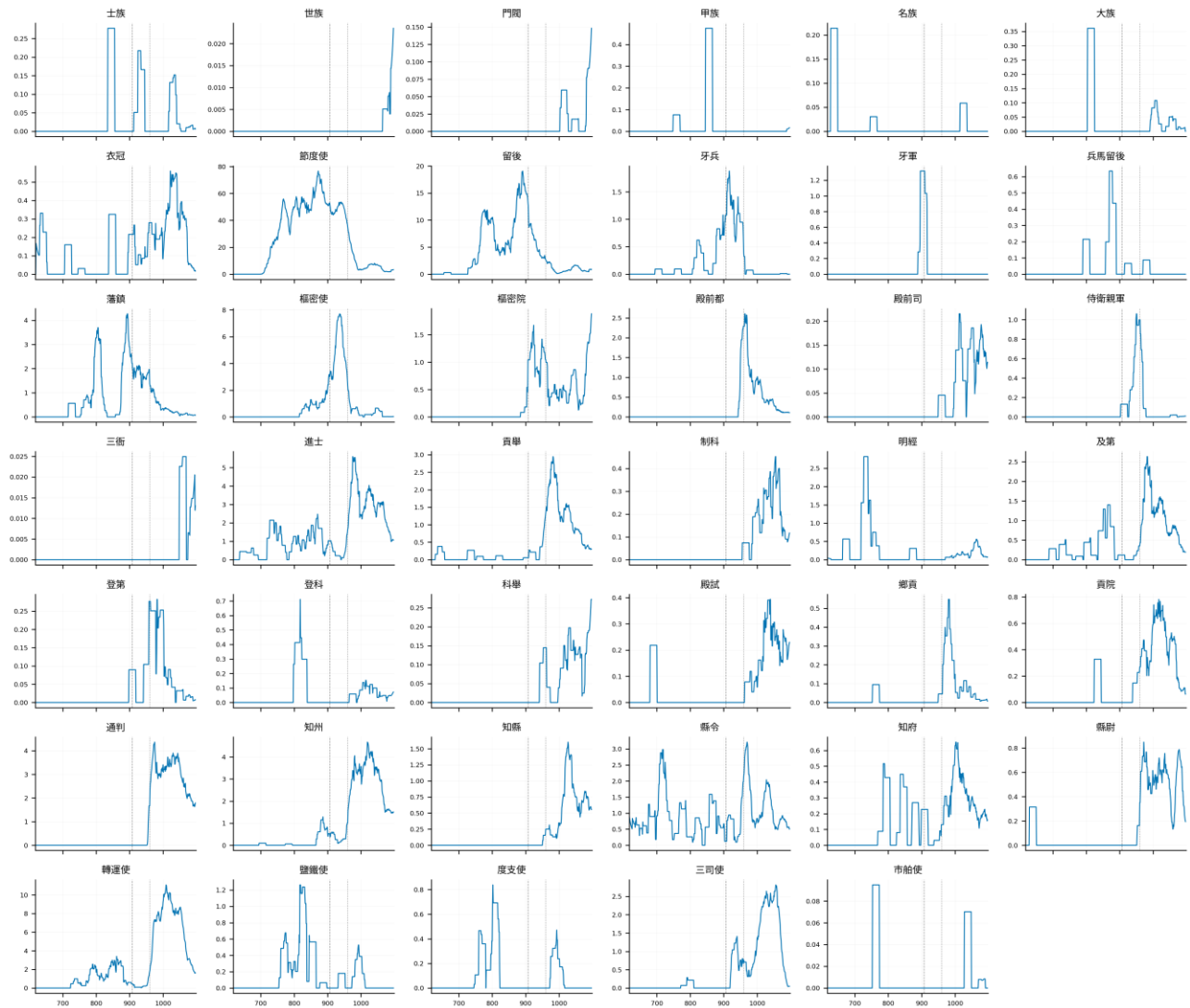


Fig. B1 Individual terms in the six-domain dictionaries, 618-1099

Notes: Annual occurrences per 10,000 characters; reference lines mark 907 and 960.

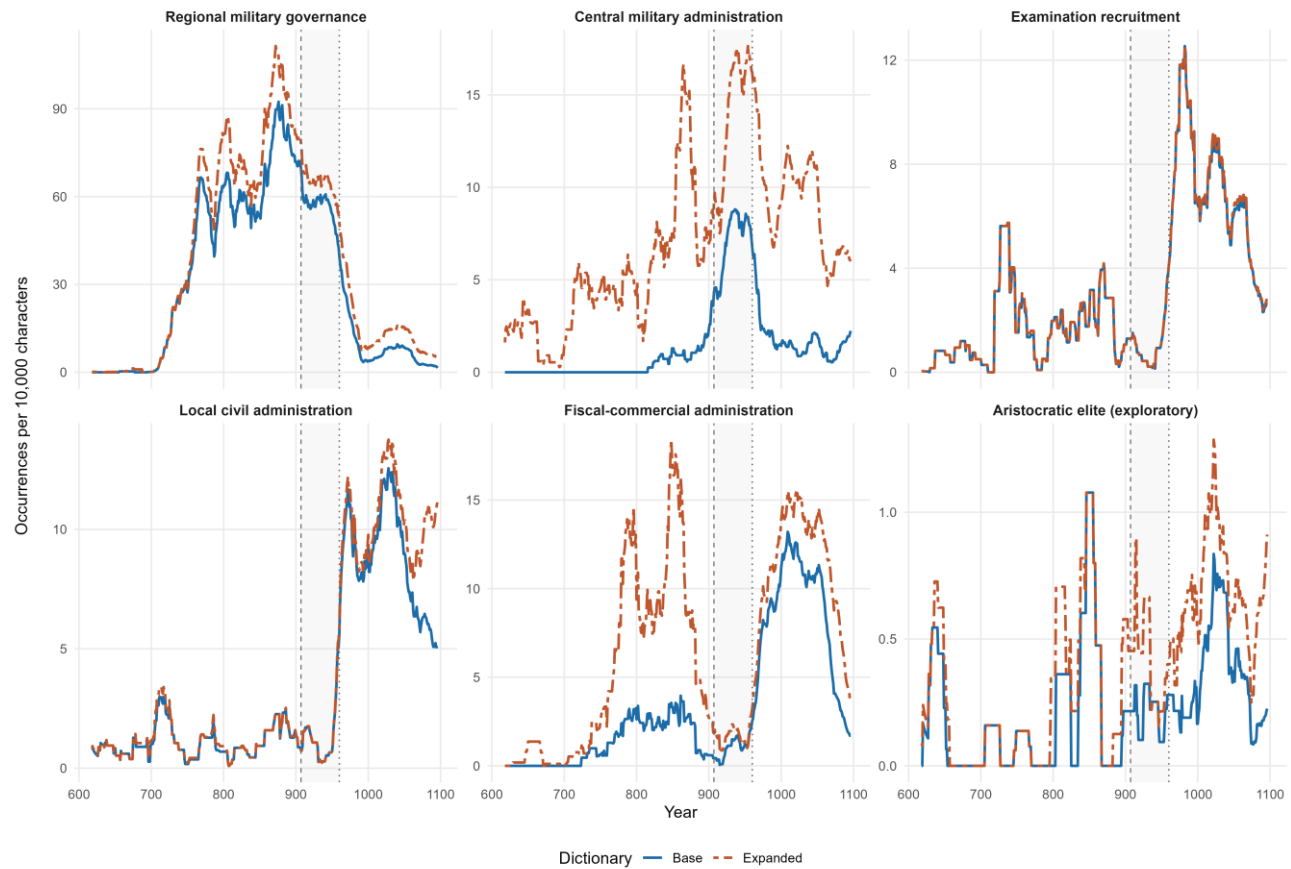


Fig. B2 Base and expanded institutional-dictionary trajectories

Notes: Solid lines show the base dictionaries and dashed lines the expanded dictionaries. Each series reports the centered 21-year moving average of occurrences per 10,000 characters. Vertical lines indicate 907 (dashed) and 960 (dotted). The expanded dictionaries include the base terms plus the additional terms listed in the note to Table B1.